Abstract

Bioinformatics is an interdisciplinary field that focuses on developing and using methods and software to understand large, complex biological data sets. Sequence alignment is a way of arranging sequences of DNA, RNA, or protein to identify regions of similarities that may explain the relationships between the sequences. There have been numerous multiple alignment tools, but one of the most commonly used one is MUSCLE, which stands for MUltiple Sequence Comparison by Log-Expectation, which uses computer science structures, such as binary trees to create a draft of the alignment. An issue that has been of interest in bioinformatics is Lepidochelys Kempii's migratory pattern as well as its recovery as a species. The original cause for the depleted L. Kempii number was egg harvesting and poaching, but now the primary threats are habitat loss, pollution, and entanglements in shrimping nets. The purpose of this experiment was to determine if there is a population similarity between the L. Kempii in Florida and the L. Kempii in Mexico. The results of this project can be used to help conservation efforts of L. Kempii. This could also reveal that there is a nesting site unknown to humans. The null hypothesis was that the two populations were genetically different, and the alternate hypothesis was that the two populations were genetically similar. The first step in the procedure for this experiment was sample collection. The samples were mitochondrial DNA of L. Kempii and were collected from Florida and Mexico. The two corresponding files (for the same sequence) were loaded in UGENE. The reversed sequences had already been reversed back to forward during the cleaning process. Then, a built-in alignment algorithm, MUSCLE, was used to align the two sequences. Finally, a final alignment with all of the samples was created. With a Fixation index value of less than .05, this experiment was able to reject the null hypothesis and conclude that the populations were genetically similar.

Introduction

Bioinformatics is an interdisciplinary field that focuses on developing and using methods and software to understand large, complex biological data sets. This field combines biology, computer science, information engineering, math, and statistics to interpret and understand biological data. Bioinformatics also refers to biological studies that use computer science and programs as a part of their methodology.

A specific part of bioinformatics is sequence alignment. Sequence alignment is a way of arranging sequences of DNA, RNA, or protein to identify regions of similarities that may explain the relationships between the sequences. In most bioinformatics programs, the aligned sequences are rows within a matrix. In sequence alignment, there are two main alignment methods, pairwise and multiple. Pairwise alignment involves finding best-matching piecewise alignments for two sequences. They are usually used when there is not a need for extreme precision, such as searching a database for sequences with a high similarity to the original sequence.

Multiple sequence alignment is an extension of pairwise alignment. Multiple alignments are used to identify similar or identical regions across a group of sequences hypothesized to be evolutionarily related. They are also used to construct phylogenetic trees by using the evolutionary relationships that can be determined by the similarity of certain regions of the sequence. There have been numerous multiple alignment tools, but one of the most commonly used one is MUSCLE, which stands for MUltiple Sequence Comparison by Log-Expectation, which uses computer science structures, such as binary trees to create a draft of the alignment. The entire algorithm has a time complexity of O(N^3), which is due to its complex 3 step algorithm which first creates an alignment and improves the alignment (Edgar, 2004).

An issue that has been of interest in bioinformatics is Lepidochelys Kempii's migratory pattern as well as its recovery as a species. L. Kempii is a critically endangered species. Currently, bioinformatics is being used to track the movement patterns of the species, especially the movement of juveniles. The study of juvenile L. Kempii is especially important to this endangered species since juveniles make up 81-87% of the species. Furthermore, computer science has greatly increased the ability to understand the genetic makeup of populations. Bioinformatics allows researchers to compare the similarity of populations and the level of cross-population mating between the two populations. This can be applied to this research as well to understand more about the genetic variation between populations.

Lepidochelys Kempii, also known as Kemp's ridley sea turtle or as the Atlantic ridley sea turtle, is a critically endangered species of sea turtle. It is also the rarest species of sea turtle. L. Kempii is the smallest of all sea turtles, weighing 79-99 lbs and having a length of 23-28 inches. L. Kempii reach sexual maturity and change from juveniles to sexually mature adults around 10-12 years (NOAA Fisheries, 2020).

L. Kempii are found in the Atlantic Ocean and the Gulf of Mexico. The turtles prefer warm waters, but have been in waters as far north as New Jersey, USA. These turtles usually migrate to the Gulf of Mexico and the Western Atlantic, in Louisiana. Mature L. Kempii are rarely found outside the Gulf of Mexico, but juveniles migrate along the East Coast of the United States (Shaver et al, 2008). The makeup for the turtles that are found in the Atlantic is the following: 2-4% are adults, 11-15% sub-adults (no longer a juvenile, but has not reached the weight necessary to sexually reproduce), and 81-87% are juveniles (Ernst et al, 2009).

The turtles nest primarily during the day from April to August. The nesting sites are mostly in Tamaulipas, Mexico, along a 16 mile stretch of beach, and Padre Island in Texas. L.

Kempii prefer to nest in areas with dunes or swamps. The female L. Kempii land in groups on the beach, which is called a mass nesting. Females usually lay around 110 eggs in one nest and they nest two-three times a year with a break of 10-12 days between each nesting (NOAA Fisheries, 2020).

The original cause for the depleted L. Kempii number was egg harvesting and poaching, but now the primary threats are habitat loss, pollution, and entanglements in shrimping nets. There have been efforts to protect L. Kempii from 1966 and the turtles have been listed under the Endangered Species Act of 1973 (Ernst et Al, 2009).

The goal of this experiment was to determine whether juvenile turtles from a nesting site in Florida were related to those in Mexico. This experiment, which was a hypothesis test, was done by comparing the genetic sequences of juvenile L. Kempii found in Florida to those found in Mexico. The null hypothesis was that there was a substantial genetic difference between the two populations. The alternate hypothesis was that there was no discernable genetic difference between the two populations. The results of this project can be used to help conservation efforts of L. Kempii. This could also reveal that there is a nesting site unknown to humans if the null hypothesis is rejected. Helping conservation efforts of L. Kempii has huge impacts on the ecosystem. If this critically endangered species was to become extinct, the delicate balance of the ecosystem could be thrown off and have devastating impacts, especially for humans. Furthermore, the same techniques discussed in this paper could be applied to other endangered species in the Atlantic ocean.

Methods and Materials

The main goal in this experiment was to determine whether there was a difference in genetic makeup between samples collected from Florida and Mexico. The first step in the procedure for this experiment was sample collection. The samples were mitochondrial DNA of L. Kempii and were collected from Florida and Mexico. There were two types of Florida samples collected, a forward sequence and a reverse sequence. These were used to create a consensus sequence that amplified DNA. When the forward sequence and the reverse sequence are combined, the consensus sequence is given a longer length and serves as a quality check if one direction is unclear. There were 102 samples collected. 21 of these samples were blood, and the rest were tissue.

After this process, the chromatograms were obtained and ranged from about 600-900 base pairs in length. In a chromatogram, a clear peak defines the base pair call. The next step was to clean the sequences by visualizing them through the chromatograms. The program used to visualize the chromatograms was FinchTV, a chromatogram viewer that helps to visualize the traces of each base. The sequences were cleaned by deleting the messy parts in the chromatograms and replacing ambiguous peak calls in the middle of the sequence with an ‘N’. The beginning and ending sections had a tendency to be “messier” than the midsection, which was of higher quality. A “messy” section (Figure 1) was defined by an unclear peak call, for example if two peaks overlapped, and it was not clear which one was higher.

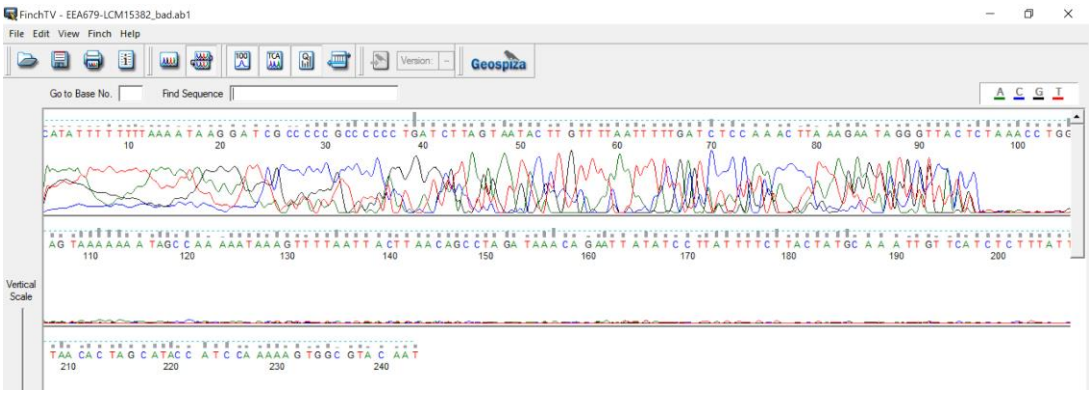

Figure 1: Messy data, found towards the end of a sequence.

There were some sequences where it was “messy” throughout the entire sequence. These sequences were marked as unusable and were excluded from the alignment. An example of a clean sequence can be found in Figure 2.

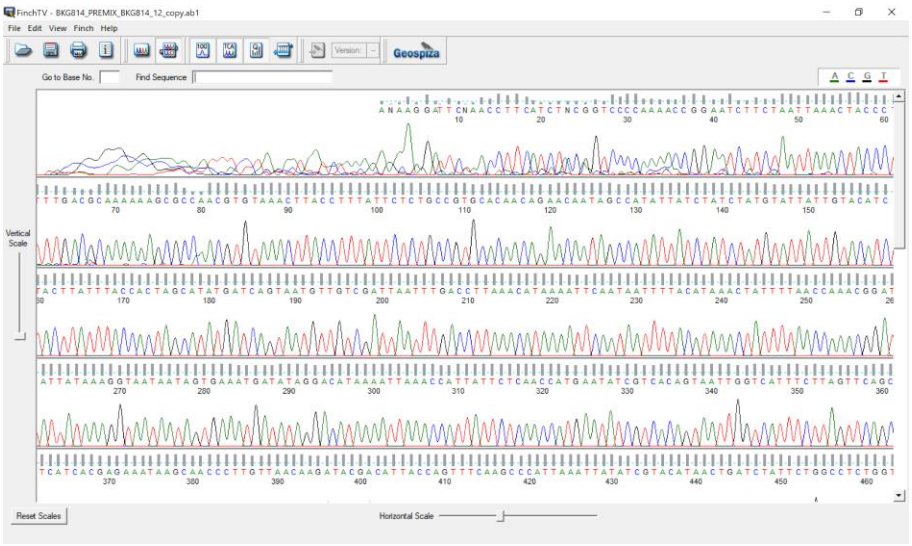

Figure 2: Clean Data

One substep of cleaning the chromatograms was changing the direction of the reverse Florida samples (H590g) while cleaning them because these would later be used to create a consensus sequence, along with the forward Florida sequences (LCM15382).

The next step was to create the consensus sequences. This was done only for the Florida sequences. However, not all Florida sequences had consensus sequences. Some were forward, while others were a consensus sequence.This was done using the program UGENE, a computer bioinformatics software that was built using C++. UGENE helps to analyse biological datasets (such as sequences, annotations, multiple alignments, phylogenetic trees, and NGS assemblies). This experiment mostly focuses on UGENE's multiple alignment properties. The two corresponding files (for the same sequence) were loaded in UGENE. The reversed sequences had already been reversed back to forward during the cleaning process. Then, a built-in alignment algorithm, MUSCLE, which stands for MUltiple Sequence Comparison by Log Expectation, was used to align the two sequences. The MUSCLE algorithm has 3 stages. The first stage is the draft progressive. During this stage, the algorithm produces a draft multiple alignment sequence, prioritizing speed over accuracy. The next step of the algorithm is called the improved progressive, which uses a binary tree to create a more accurate alignment. A binary tree is a simple decision making data structure used in computer science to help a computer search faster. At the end of the first two steps, the algorithm has a time complexity of O(N^2). The final step of the algorithm is called the refinement stage. The refinement stage corrects any mistakes made during the previous stages. After the final stage has finished, the algorithm has a time complexity of O(N^3). Figure 3 shows the MUSCLE algorithm in detail below.

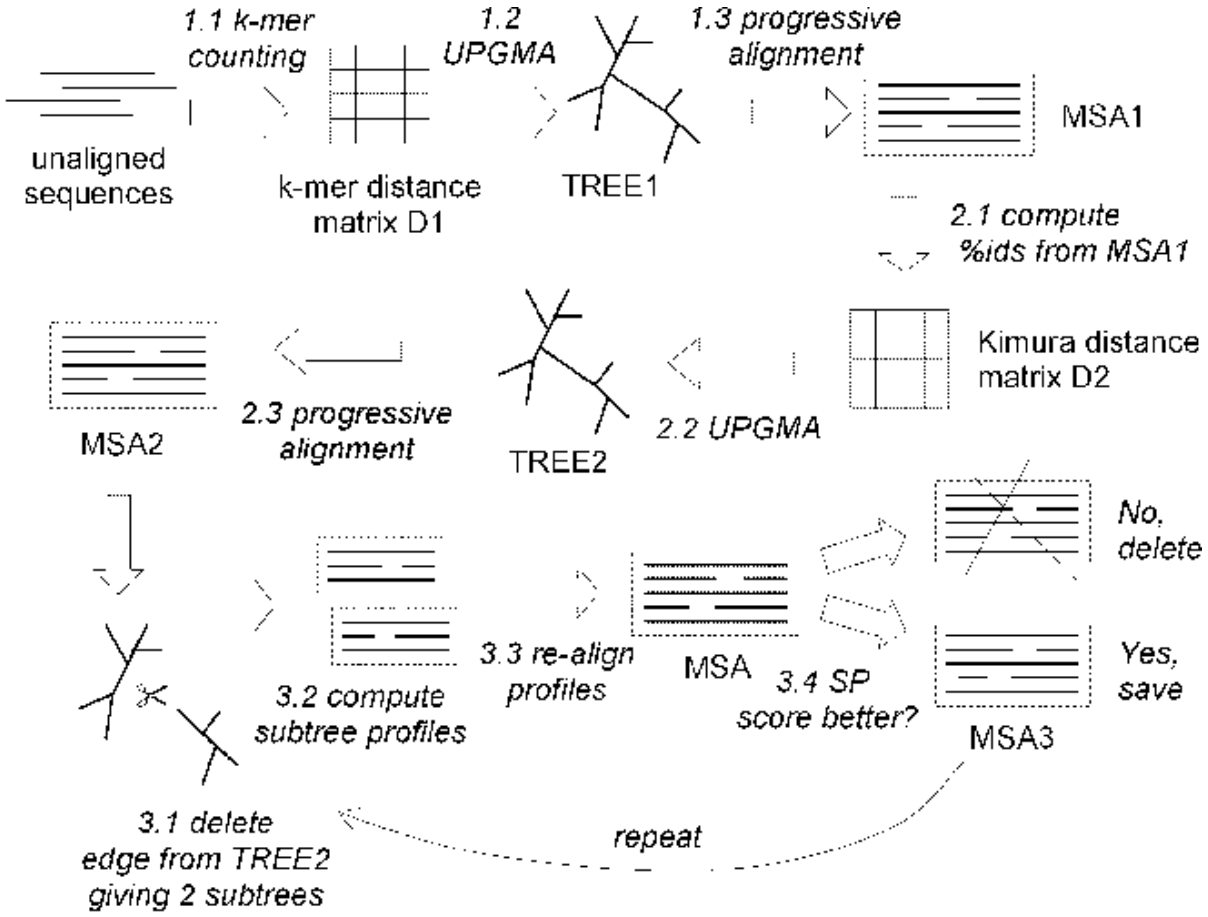

Figure 3: Complete detail of MUSCLE Algorithm. This shows the way that the computer makes the decisions to come up with the ideal alignment using different distance measures to determine how similar two alignments are.

Variations between the sequences in the alignment were checked and resolved by verifying the correct base pair call through chromatograms or by marking it as an "N," an ambiguous call. These consensus sequences were used in the final alignment in place of the forward and reverse sequences that they originated from.

Finally, a final alignment with all of the samples was created. First, all of the samples (the forward Mexico samples, the forward Florida samples, and the consensus Florida samples) were loaded into the program UGENE. Then, the sequences were aligned through MUSCLE. These sequences went through the same algorithm as before. However, this step took longer

because there were more sequences to align. This is because the time complexity, which is how long it takes for a computer to finish a task based on the size of the dataset, was O(N^3). This is highly inefficient, however MUSCLE was the only alignment algorithm in UGENE. Variations between the sequences were checked by re-examining the chromatograms. This was done in order to make sure that the variations were actual genetic variations and not errors between the sequences. An example of an error could be that the chromatogram displays the wrong base pair name. Another error could be two peaks of the same size being in the same area, making it unclear which is the correct base pair. This can be solved by either zooming in on the chromatogram or by declaring it as an ambiguous call and replacing it with “N.”

Results

Figure 4 below shows the haplotype network, which demonstrates the different relationships among the genes observed in the dataset. This is a rooted node, which would mean that the big dot (BKGB11_Premix_BKGB11_3_copy) is the first sequence in the dataset and is used as the default for the sequence that the other sequences connect to.

Figure 5 shows the Fst, which stands for the fixation index. This was acquired from using the mitochondrial DNA. The formula for Fixation index is found in Figure 6. The fixation index is a statistic used in bioinformatics and biology to determine the population differentiation due to genetic structure. It is a special case of Wright's F-Statistic. It falls between 0-1, with 0 being no variation and 1 being highly divergent. With a fixation index of below 0, we can round it up to 0. This is due to computer rounding issues. Since the fixation index is 0, we can conclude that the two populations compared (Florida vs Mexico) are both highly similar and have small or insignificant differences and that there is constant interbreeding between both the populations. There are also other values such as the GammaSt (used to calculate gene flow), Nst (a gene found in living creatures) , Dxy (Nucleotide difference), and Da(Daltons) that are generated by MUSCLE, but those can be ignored because they are not pertinent to finding the similarity between two populations.

Figure 4: Haplotype network for the sequences.

POPULATION 1	POPULATION 2	GammaSt	Nst	Fst	Dxy	Da
Florida	Mexico	0.02187	-0.02670	-0.02674	0.00106	-0.00003

Figure 5: Statistics for the two populations' comparison.

$$F_{ST} = \frac{\bar{p}(1-\bar{p}) - \sum c_i p_i (1-p_i)}{\bar{p}(1-\bar{p})} = \frac{\bar{p}(1-\bar{p}) - \overline{p(1-p)}}{\bar{p}(1-\bar{p})}$$

Figure 6: Formula for the Fixation index. Diversity in this formula is calculated using the probability that two randomly selected alleles are similar, i.e. when 2p(1-p) with p being the frequency of the allele in the populations. This helps to explain the amount of genetic variation due to the population structure. The first part of the formula in the numerator (before the second negative sign) is for population 1 (in this case, Florida) and the second part (after the second negative sign) is for population 2. The difference is divided by the population 1 to get the Fixation Index.

Discussion and Conclusions

The purpose of this experiment was to determine if there is a population similarity between the L. Kempii in Florida and the L. Kempii in Mexico. The null hypothesis was that the two populations were genetically different, and the alternate hypothesis was that the two populations were genetically similar. With a Fixation index value of less than .05, this experiment was able to reject the null hypothesis and conclude that the populations were genetically similar.

It is important to track the movement of L. Kempii because it is a critically endangered species. Tracking the movement of juveniles is especially important, since they make up well over half of the population. This could help conservation efforts by closely monitoring the areas that L. Kempii travel and nest so that stricter rules and regulations can be set in regards to fishing and other water activities. It is important to conserve this species because if the conservation efforts are successful, scientists are able to apply these same methods to help other species.

For further study and improvements to the experiment, the size of the sample could be increased. This would make sure that the experiment accounts for more of the variations that are present in the L. Kempii genome. Also, there could be a more diverse sample that was collected on the beaches of Mexico. This more diverse sample could lead to understanding exactly where the group of turtles from Florida come back to land and lay eggs in Mexico. Furthermore, a different and more efficient algorithm could have been used. While the MUSCLE algorithm is extremely accurate, it is so at a cost of speed and computer space. A time complexity of O(N^3) means that the algorithm went through the data 3 times, which is fine for smaller datasets such as the ones that were used now, but for larger datasets could mean that it would take days to process.

Works Cited

Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113 (2004). https://doi.org/10.1186/1471-2105-5-113

Fisheries, NOAA. "Kemp's Ridley Turtle." *NOAA*, 2020, www.fisheries.noaa.gov/species/kemps-ridley-turtle#:~:text=NOAA%20Fisheries%20is%20committed%20to,treaties%2C%20to%20protect%20sea%20turtles.

"L. Kempii." *Turtles of the United States and Canada*, by Carl H. Ernst and Jeffrey E. Lovich, The Johns Hopkins University Press, 2009, p. 827.

Shaver, Donna & Rubio, Cynthia. (2008). Post-nesting movement of wild and head-started Kemp's ridley sea turtles Lepidochelys kempii in the Gulf of Mexico. Endangered Species Research. 4. 43-55. 10.3354/esr00061.